NDB-JEPA: LABEL-EFFICIENT GRASP DETECTION FROM FROZEN I-JEPA FEATURES

Anonymous authors
Paper under double-blind review

ABSTRACT

As robots become increasingly prevalent in real-world settings, their perception models must become label-efficient and robust to distributional shifts not represented in their labeled training data which poses a considerable challenge for end-to-end grasp detection models typically trained on limited or synthetic datasets. We introduce NDB-JEPA, a label-efficient grasp point detection perception model centered around I-JEPA. Our proposed model takes advantage of the robust, generalized feature extraction of I-JEPA, a self-supervised joint-embedding predictive foundation model. Our model trains lightweight grasp heads on top of the frozen image patch outputs of a 632M-parameter I-JEPA encoder, predicting grasp confidence, orientation, and width per embedded patch. We ask whether these frozen features can replace end-to-end training for the grasping problem. On the Cornell Grasping Dataset, NDB-JEPA reaches 86.9% Top-1 rectangle success, within 1.5 points of an end-to-end GR-ConvNet baseline, and among large pretrained encoders, I-JEPA features outperform MAE, supervised ViT, and DINOv2 alternatives. The frozen representation is also markedly more label-efficient and robust: with only 10% of the training data, NDB-JEPA attains roughly 74% success versus 40% for GR-ConvNet trained from scratch, and it retains about 94% of its clean accuracy under severe darkening while the end-to-end baseline collapses to 1.1%. While end-to-end models retain a small edge in clean accuracy, our results indicate that frozen joint-embedding predictive representations offer a data-efficient, robust basis for grasp perception.

1 INTRODUCTION

Grasp point detection is a core perception problem for robotic arm manipulation tasks. Given an RGB or RGB-Depth (RGB-D) image of a scene with a target object, an algorithm predicts an end-effector grasp configuration that can reliably grab the target object (Lenz et al., 2015). A grasp point is made up of a grasp center, grasp angle, and grasp width. Generalizability is key to these models being used in the real-world, which is the main goal of our model (Dosovitskiy et al., 2017). A model that can work in a variety of new or unusual environments that one might come across in the wild is key to robotics becoming more commonplace. However, making a labeled grasping dataset covering all these environments is infeasible, so we believe that robotic perception models should be based on large foundational models and then fine-tuned to their specific problems. We architected our solution to use the I-JEPA model (Assran et al., 2023), a foundational vision model trained in a self-supervised manner, in order to improve the generalizability of our model. Specifically, using the vision transformer encoder component of I-JEPA as a frozen feature extractor. We test two hypotheses: (i) that I-JEPA's latent-predictive pretraining yields patch representations from which accurate grasps can be decoded by a lightweight head, without any fine-tuning of the backbone; and (ii) that these frozen representations are more label-efficient and more robust to appearance shift than an equivalent grasp network trained end-to-end from scratch.

Grasp datasets like Cornell (Jiang et al., 2011), Jacquard (Depierre et al., 2018), or Dex-Net (Mahler et al., 2017) are either synthetic or simple images of single objects on a flat surface. Good performance on these datasets may not correspond to good performance for a perception model outdoors trying to pick up an object in a cluttered environment or with different lighting (Geirhos et al., 2019). If our I-JEPA architecture performs comparably while generalizing well, then there is evidence that

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

foundation based perception models are key for getting object manipulation to work in dynamic, messy real world environments.

However, whether frozen foundation-model features actually help grasp detection and which pretraining objective yields the most useful features has not been established in a controlled way: a representation can look strong simply because its encoder is larger or because evaluation never leaves the clean training distribution. We therefore structure this work as a controlled study rather than a single system. Holding the grasp head, data splits, and evaluation fixed, we vary only the pretrained encoder, comparing latent-predictive I-JEPA features against reconstructive (MAE) (He et al., 2022), supervised (Dosovitskiy et al., 2021), and self-distilled (DINOv2) (Oquab et al., 2024) alternatives at matched scale. We then measure how each representation behaves as labeled data shrinks to 5% of the training set, and how the resulting detector degrades under Cornell-C, a suite of controlled image corruptions, relative to an end-to-end baseline. Our contributions are as follows:

- We introduce NDB-JEPA, which trains lightweight confidence and geometry heads on frozen I-JEPA patch embeddings, reaching 86.9% Top-1 rectangle success on the Cornell Grasping Dataset, within 1.5 points of an end-to-end GR-ConvNet baseline.
- We present a controlled comparison of frozen pretraining paradigms for grasp detection at matched encoder scale, showing that I-JEPA provides the strongest features and that additional capacity alone does not close the gap.
- We show that frozen predictive features are markedly more label-efficient: with only 10% of the training labels, NDB-JEPA reaches approximately 74% success versus approximately 40% for GR-ConvNet trained from scratch.
- We construct Cornell-C, a corruption suite for Cornell, and show that frozen features retain roughly 94% of clean performance under severe darkening and 88% under occlusion where the end-to-end baseline collapses; a modality-matched RGB-only baseline reveals that the baseline's remaining advantages under noise and clutter are largely attributable to its uncorrupted depth channel.

2 Related Work

Grasp detection. The current practices train models end-to-end on large synthetic databases of objects with simulation labeled grasp points. Current state of the art models include GQ-CNN, GG-CNN, and GR-ConvNet (Mahler et al., 2017; Morrison et al., 2020; Kumra et al., 2020). These models see great performance on their limited datasets; however, aside from data augmentation, little effort is put into generalizing these models. Some models like GQ-CNN saw their biggest improvements with dataset changes; they generated a giant synthetic dataset of grasps called Dex-Net which improved the performance of their model by 4% (Mahler et al., 2017). However, these models are trained on synthetic datasets, leading to the sim-to-real problem which is a big hurdle in robotics (Tobin et al., 2017). The real world performance of these models often suffers (Bousmalis et al., 2018), and the current direction of research is in improving results further instead of generalizing.

Self-supervised visual representations. Vision transformers (Dosovitskiy et al., 2021) are the standard backbone for large-scale visual pretraining, and the major self-supervised paradigms differ chiefly in their objective: DINOv2 learns by discriminative self-distillation (Oquab et al., 2024), MAE reconstructs masked pixels (He et al., 2022), and I-JEPA predicts the latent representations of masked regions rather than their pixels (Assran et al., 2023). Representation quality in this literature is typically assessed by freezing the encoder and training small probes; we extend that protocol from classification to dense grasp prediction.

Pretrained representations for robotics. A growing body of work asks whether general visual pretraining transfers to embodied tasks. R3M pretrains on human video for manipulation (Nair et al., 2022), masked visual pretraining has been applied to real-robot control (Radosavovic et al., 2023), and VC-1 benchmarks frozen visual encoders across embodied-AI tasks (Majumdar et al., 2023). These efforts largely evaluate representations through downstream policy learning; we instead evaluate them on grasp rectangle prediction, a dense perception task with a standardized metric, which

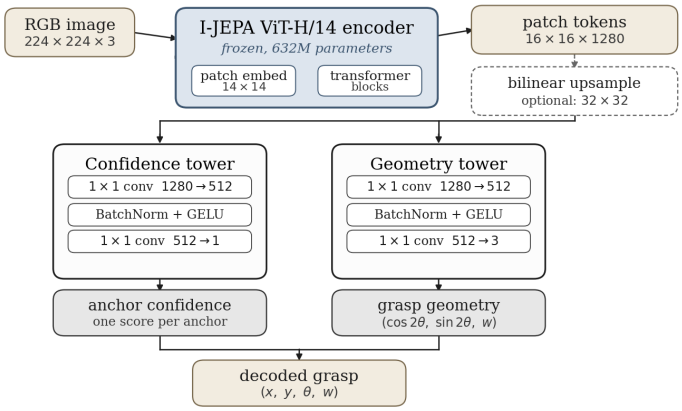


Figure 1: Architecture overview: a frozen I-JEPA encoder produces patch embeddings and a two-tower head predicts a grasp confidence map and the corresponding grasp geometry at every anchor.

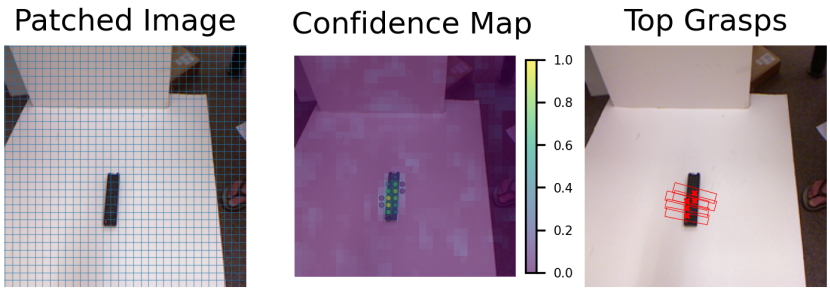


Figure 2: Left: the upscaled 32×32 grid of I-JEPA patch embeddings Middle: confidence scores for each patch Right: the top predicted grasp rectangles overlaid on the input.

isolates representation quality from policy-optimization effects. Transformer architectures have also been trained end-to-end for grasp detection (Wang et al., 2022), but without leveraging frozen foundation-model features.

Robustness to distribution shift. Vision models degrade under common corruptions such as noise, blur, and lighting changes (Hendrycks & Dietterich, 2019), and ImageNet-trained CNNs are known to over-rely on texture (Geirhos et al., 2019). In robotics this manifests as the sim-to-real gap, commonly mitigated with domain randomization (Tobin et al., 2017). Our Cornell-C suite follows the design of ImageNet-C at the grasp-dataset level, enabling a controlled comparison of frozen and end-to-end detectors under appearance shift.

3 Methodology

We use a pretrained I-JEPA model as a frozen feature extractor and train lightweight grasp-prediction heads on its patch embeddings. Our system architecture is shown in Figure 1.

3.1 Frozen I-JEPA Backbone

We extract patch-level embeddings from a pretrained I-JEPA ViT, accessed through the Hugging Face SDK (Meta AI, 2023). Specifically, we use the ViT-H/14 I-JEPA model pretrained on ImageNet-1K. Choosing the smaller of the ImageNet pretraining sets reduces our compute requirements, and the 14-pixel patches yield a denser grid of anchor points (potential grasp centers) within each image. All backbone parameters remain frozen; only the grasp heads are trained.

3.2 Grasp Heads

A lightweight two-tower head operates on the grid of patch embeddings. The confidence tower predicts, for every patch, the likelihood that a grasp center lies within it; we call these locations grasp anchors, and the resulting map forms an anchor confidence matrix; Figure 2 shows an example. The geometry tower predicts the grasp orientation together with the gripper width. Both towers are implemented as 1×1 convolutions, so the same head is shared across all spatial locations. The exact topology and training objective are given in the Experiments section.

3.3 Architecture Rationale and Motivation

We expected this approach to succeed because JEPA-style models have been shown to provide strong foundation features across a variety of tasks (Assran et al., 2023; LeCun, 2022). Anchoring predictions at patch centers also yields a considerably simpler architecture than detection models such as DETR (Carion et al., 2020), which require a dedicated transformer decoder.

4 Experiments

4.1 Setup and Dataset

All experiments were run on a single NVIDIA GeForce RTX 3060 with 12 GB using Windows 10, PyTorch 2.11, the Hugging Face Transformers library 5.5, CUDA 13.0, and Python 3.14. Our grasp-head and frozen backbone 20-epoch configuration trains in approximately 15 minutes for the ViT-H/14 encoders and the complete set of runs reported in this paper required roughly 22 GPU-hours.

All experiments use the Cornell Grasping Dataset: real RGB-D images of physical objects, each annotated with multiple hand-labeled grasp rectangles encoding a grasp center, orientation, and gripper width. Cornell is publicly available and uses real visual data, making it a realistic benchmark, but it is small, and its human labels may be biased toward grasps that are natural for people rather than optimal for a parallel-plate gripper (Jiang et al., 2011; Lenz et al., 2015). Models trained only on Cornell should therefore not be assumed to generalize to unseen object categories or cluttered scenes, a limitation that our corruption suite probes directly.

4.2 Split and Evaluation Metric

We partition the data with a 80:10:10 image-wise split. The split contains 708, 88, and 89 images. Cornell's 885 images show 240 objects, about 3.7 views each, so most test objects also appear in the training set from other viewpoints: the test set measures generalization to new views of seen objects rather than to unseen objects. A planar parallel-jaw grasp is represented as follows: $g = (x, y, \theta, w)$, where (x, y) is the image-plane grasp center, θ is the in-plane gripper orientation, and w is the gripper opening. Because a parallel-jaw grasp is invariant to a rotation of π, the angular error is computed as

$$\Delta_\theta(\widehat{\theta}, \theta) = \min\left(\delta,\, \pi - \delta\right), \qquad \delta = \left|\widehat{\theta} - \theta\right| \bmod \pi. \tag{1}$$

We report Top-1 rectangle success (Jiang et al., 2011; Lenz et al., 2015). For test image i and IoU threshold τ, a predicted grasp $\widehat{g}_i$ is considered successful when at least one ground-truth grasp $g \in \mathcal{G}_i$ satisfies

$$s_i = \mathbb{I}\Big[\exists g \in \mathcal{G}_i : \mathrm{IoU}\left(R(\widehat{g}_i), R(g)\right) > \tau \wedge\ \Delta_\theta\left(\widehat{\theta}_i, \theta_g\right) < 30^\circ\Big]. \tag{2}$$

The reported success rate is $\mathrm{Success@1} = \frac{100}{N}\sum_{i=1}^{N} s_i$.

All methods are decoded and evaluated using the same implementation. The ground-truth rectangle $R(g)$ is taken directly from the four annotated corners of the Cornell grasp label, so no ground-truth geometry is synthesized. The grasp head regresses a center, an angle and a gripper opening but

not a jaw length, so the jaw extent of the *predicted* rectangle is set to twice the predicted opening, following the convention that the two rectangle sides differ by a factor of two (Morrison et al., 2020). We report success at the standard $\mathrm{IoU} > 0.25$ operating point used throughout the grasp-detection literature (Jiang et al., 2011; Lenz et al., 2015), and additionally at the stricter $\mathrm{IoU} > 0.50$; the latter should be interpreted relative to the discretization ceiling quantified in Appendix E, since the 16×16 anchor grid caps attainable strict-threshold success at 64.0%. Appendix C explains why absolute values under this protocol are lower than published Cornell results.

4.3 Frozen-Backbone Comparison

To compare predictive, reconstructive, supervised, and discriminative self-supervised pretraining without confounding the comparison with model capacity, we evaluate four large pretrained encoders: I-JEPA ViT-H/14, MAE ViT-H/14, supervised ViT-H/14, and DINOv2 ViT-g/14. The first three are tightly matched in parameter count (approximately 632M, 631M, and 632M) and patch size. DINOv2 ViT-g/14 is the closest available large DINOv2 checkpoint but is larger at approximately 1.1B parameters, so we describe this as a large-scale rather than a strictly parameter-matched comparison; as reported below, its extra capacity does not translate into stronger grasp features.

4.4 Common Grasp Head and Training Protocol

All backbone parameters remain frozen throughout training. Each RGB image is resized to 224×224 without center cropping. Non-spatial prefix tokens, including class and register tokens where applicable, are removed before grasp prediction. The remaining patch tokens are reshaped into their native spatial grid and bilinearly resampled to a common 16×16 grid.

The same two-tower grasp-head topology is applied to each frozen backbone. The confidence tower predicts a grasp score at every spatial location, which are the grasp anchors, while the geometry tower predicts $(\cos(2\theta), \sin(2\theta), w)$. The doubled-angle representation handles the π-periodicity of a parallel-jaw grasp. Each tower consists of a 1×1 input projection, batch normalization, GELU activation, spatial dropout, and a final 1×1 prediction layer. Each tower uses a hidden dimension of 512, with the input projection sized to the backbone embedding dimension (1280 for the ViT-H/14 encoders and 1536 for DINOv2 ViT-g/14). All configurations train with a learning rate of 5×10^{-4}, at batch size 32 for the backbone comparison, low-shot, and corruption experiments. Backbone weights are held in evaluation mode with gradients disabled throughout, so the trainable parameters are confined to the two towers.

The training objective is $\mathcal{L} = \mathcal{L}_{\mathrm{conf}} + 0.1\mathcal{L}_{\mathrm{geom}}$, where $\mathcal{L}_{\mathrm{conf}}$ is mean squared error over the confidence map. Geometry loss is evaluated only at positive grasp-anchor locations:

$$\mathcal{L}_{\mathrm{geom}} = \frac{1}{3\sum_{u,v} M_{uv}} \sum_{u,v} M_{uv} \left\| \widehat{\mathbf{z}}_{uv} - \mathbf{z}_{uv} \right\|_2^2, \tag{3}$$

where $M_{uv} = \mathbb{I}\left[Y^q_{uv} > 0.5\right]$ and $\mathbf{z}_{uv} = (\cos(2\theta), \sin(2\theta), w)$.

All configurations within a comparison use the same data split, loss, checkpoint-selection criterion, and evaluation implementation. The checkpoint with the minimum validation loss is evaluated on the held-out test set. Each configuration is repeated using seeds $\{0, 1, 2\}$, and results are reported as the mean and sample standard deviation. Appendix A reports the design ablations and the hyperparameter search, and Appendix B their validation-split re-run.

4.5 Low-Shot Sample Efficiency

We evaluate sample efficiency using 5%, 10%, 25%, 50%, and 100% of the data. A single deterministic permutation of the training indices is produced using the fixed data seed. Each fraction uses a prefix of this permutation, which makes the subsets nested: every example used by a smaller fraction is also contained in all larger fractions. The validation and test sets remain fixed across fractions, models, and seeds.

Table 1: Frozen-backbone Top-1 rectangle success on Cornell under the canonical metric, at $\mathrm{IoU} > 0.25$ and $\mathrm{IoU} > 0.50$. Values are mean $\pm$ sample standard deviation over three seeds. The first three models are matched at approximately 632M parameters, while DINOv2 ViT-g/14 is larger. The final row gives the ceiling imposed by the anchor grid, which binds only at the stricter threshold.

Backbone	Params	@0.25	@0.50
I-JEPA ViT-H/14	632M	$\mathbf{86.9 \pm 1.3}$	$\mathbf{34.8 \pm 1.1}$
MAE ViT-H/14	631M	83.5 ± 1.3	33.0 ± 2.3
Supervised ViT-H/14	632M	79.0 ± 1.3	27.7 ± 3.6
DINOv2 ViT-g/14	1.1B	73.4 ± 1.7	25.1 ± 3.4
Grid oracle ceiling (16×16)		100.0	64.0

The low-shot experiment includes four frozen-backbone systems and one grasp-specific model trained from scratch; I-JEPA, MAE, supervised ViT, DINOv2, and end-to-end GR-ConvNet trained from random initialization.

For the frozen-feature methods, only the grasp head is optimized. For GR-ConvNet, all convolutional and residual layers are trained.

All models use the same held-out images and the same Top-1 rectangle metric. The frozen and end-to-end models also use different batch sizes because of their different memory requirements.

4.6 Out-of-Distribution Corruption Evaluation

We construct a Cornell-C test suite to measure robustness to appearance changes absent from the clean training distribution. All models are trained only on clean Cornell images. At evaluation time, the held-out RGB images are transformed by one of seven corruptions: darkening (brightness factor 0.4), brightening (factor 1.8), contrast reduction (factor 0.4), additive Gaussian RGB noise (standard deviation 30 on the $[0, 255]$ intensity scale), Gaussian blur (radius 4), occlusion by four randomly positioned 70×70 black regions, and synthetic clutter produced by pasting three object crops taken only from Cornell training images. Because the pasted objects carry no grasp labels and the detector is not told which object to grasp, a physically valid grasp on a pasted object counts as a failure; clutter therefore measures whether a detector keeps selecting the original object, a target-selection behavior rather than robustness to appearance change.

Corruption parameters are deterministically seeded from the image path, ensuring that every model and experimental seed observes the same corrupted version of a test image. Corruptions are applied at the original image resolution before model-specific preprocessing.

5 Results

5.1 Backbone Representation Quality

Table 1 reports the frozen-backbone comparison.

I-JEPA provides the strongest frozen representation at $86.9 \pm 1.3\%$. The gap to MAE is 3.4 percentage points, which is comparable to the seed-to-seed variation of the two models, so the predictive and reconstructive objectives are close at this scale. The gap to supervised ViT-H/14 is larger at 7.9 points, with standard deviations of only 1.3 points on both models, indicating that at matched capacity and batch sizes (size 32) the two self-supervised objectives remain separated from supervised pretraining.

Model size alone does not explain the ordering. DINOv2 ViT-g/14 contains approximately 1.1B parameters, substantially more than the other encoders, yet remains 13.5 points below I-JEPA and 5.6 points below the much smaller supervised ViT-H/14. Additional capacity therefore does not make all frozen features equally effective for a dense grasp head; the pretraining objective governs whether the resulting patch embeddings are spatially informative enough for grasp localization.

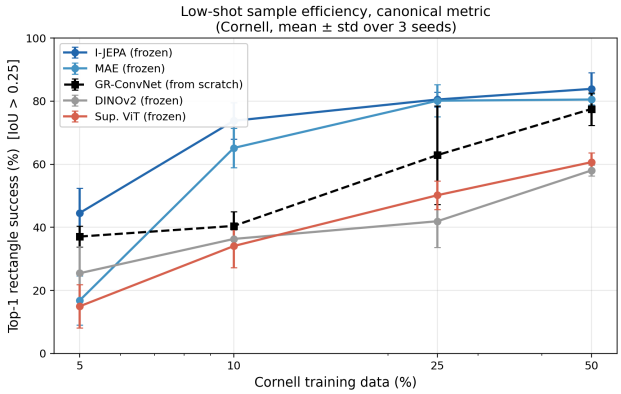


Figure 3: Low-shot sample efficiency on the Cornell Grasping Dataset. I-JEPA and MAE provide the strongest results in the 10–25% data regime. In contrast, the tested supervised ViT and DINOv2 representations remain below the from-scratch GR-ConvNet baseline throughout the plotted fractions.

The same ordering holds at the stricter IoU > 0.50 threshold, where I-JEPA reaches $34.8 \pm 1.1\%$ against a grid-imposed ceiling of 64.0%. That all four backbones fall well short of this ceiling indicates that the residual error at the strict threshold is dominated by localization quality rather than by the anchor grid alone. Appendix D offers a hypothesis for this ordering.

5.2 Low-Shot Sample Efficiency

Figure 3 presents Top-1 grasp success as the Cornell training fraction is reduced from 50% to 5%. The five models exhibit three distinct sample-efficiency patterns.

I-JEPA is strongest in the most label-restricted regime. At 5% of the Cornell training data it reaches approximately 45% success, compared with approximately 37% for GR-ConvNet, 25% for DINOv2, 17% for MAE, and 15% for supervised ViT; MAE degrades sharply at this fraction, so its advantage over the from-scratch baseline is not maintained in the extreme low-label regime. The clearest separation occurs at 10% of the data, where I-JEPA reaches approximately 74% and MAE approximately 65%, against roughly 40% for GR-ConvNet and 36% and 34% for DINOv2 and supervised ViT, placing I-JEPA about 33 points above the from-scratch baseline.

The advantage narrows as labels become plentiful. At 25% of the data I-JEPA and MAE both reach approximately 80% success against GR-ConvNet's 63%; by 50% the three are within a few points of one another. The frozen predictive and reconstructive representations therefore approach their 50%-data performance using only a quarter of the training partition.

Freezing a pretrained encoder is not sufficient by itself, however. Supervised ViT and DINOv2 remain below from-scratch GR-ConvNet at every plotted fraction, by approximately 13–21 points at 25% data, so the compatibility of a backbone's spatial features with dense grasp prediction matters more than the fact that it was pretrained.

5.3 Robustness to Cornell-C Corruptions

Table 2 compares the frozen I-JEPA system with end-to-end GR-ConvNet under clean and corrupted Cornell inputs. On clean images, with batch sizes of 8 for GR-ConvNet and 32 for NDB-JEPA, GR-ConvNet obtains approximately 88.4% rectangle success, compared with approximately 86.9% for NDB-JEPA. The clean difference is small relative to the changes produced by several of the corruptions.

The largest difference occurs under severe darkening, where NDB-JEPA retains approximately 81.3% success, about 94% of its clean performance, while GR-ConvNet falls to approximately 1.1%. NDB-JEPA is also ahead under brightening (78.3% versus 67.4%) and under occlusion, where it retains roughly 88% of clean performance against GR-ConvNet's 53% and shows far less

Table 2: Cornell-C Top-1 rectangle success (%), mean ± standard deviation over three seeds. The middle column is a GR-ConvNet trained on RGB only, which removes the input-modality confound: the corruptions perturb every channel the model sees, exactly as they do for NDB-JEPA. Bold marks the best modality-matched result, i.e. NDB-JEPA versus GR-ConvNet (RGB); the RGB-D column is reported for reference and is not part of that comparison.

	NDB-JEPA	GR-ConvNet	
Corruption	RGB	RGB	RGB-D
Clean	$\mathbf{86.9 \pm 1.3}$	82.0 ± 5.8	88.4 ± 3.6
Dark	$\mathbf{81.3 \pm 3.4}$	8.2 ± 14.3	1.1 ± 1.9
Bright	$\mathbf{78.3 \pm 2.8}$	61.4 ± 5.7	67.4 ± 10.0
Low contrast	$\mathbf{80.1 \pm 5.2}$	77.9 ± 9.2	83.9 ± 4.5
Noise	79.0 ± 3.2	$\mathbf{82.4 \pm 6.4}$	88.4 ± 2.8
Blur	$\mathbf{85.0 \pm 2.8}$	82.0 ± 7.8	85.4 ± 3.9
Occlusion	$\mathbf{76.0 \pm 3.4}$	35.2 ± 28.8	46.8 ± 11.5
Clutter	$\mathbf{24.3 \pm 1.3}$	18.7 ± 3.9	29.6 ± 3.4

seed variation (standard deviations of approximately 3 versus 12 percentage points). The ordering reverses under additive Gaussian noise, where GR-ConvNet stays near its clean result at approximately 88.4% against NDB-JEPA's 79.0%, and it is slightly ahead under low contrast. Synthetic clutter is the hardest condition for both systems (approximately 24.3% and 29.6%, or roughly 28% and 33% of their clean performance) so neither reliably preserves the target grasp among pasted distractors.

Because NDB-JEPA consumes RGB while the default GR-ConvNet consumes RGB-D, the corruptions above perturb every input channel of the former but leave the depth channel of the latter untouched. To remove this confound we retrain GR-ConvNet on RGB only, so that both systems observe exactly the same corrupted signal. The RGB-only baseline is slightly weaker on clean images (82.0% versus 88.4%), which is the expected cost of discarding depth, so it is compared against NDB-JEPA rather than against its own RGB-D variant.

The modality-matched comparison strengthens two of our conclusions, overturns one, and weakens another. First, the collapse under darkening is not a depth artifact: the RGB-only GR-ConvNet still fails almost completely ($8.2 \pm 14.3\%$, with two of three seeds near zero) while NDB-JEPA retains 81.3%. Second, the occlusion advantage persists and widens (76.0% versus 35.2%). Third, the apparent GR-ConvNet advantage under Gaussian noise shrinks from 9.4 to 3.4 points once depth is removed (82.4% versus NDB-JEPA's 79.0%), within the RGB-only baseline's seed spread, indicating that it was largely driven by the uncorrupted depth channel rather than by end-to-end training. Fourth, the clutter ordering reverses: NDB-JEPA (24.3%) outperforms the modality-matched baseline (18.7%), so the earlier clutter result was likewise attributable to depth.

Under matched input modality, then, NDB-JEPA equals or exceeds the end-to-end baseline on every corruption tested except Gaussian noise. We nonetheless temper this conclusion in two ways. The RGB-only baseline exhibits very large seed variance under darkening and occlusion (standard deviations of 14.3 and 28.8 points), so its degraded conditions are unstable rather than uniformly poor. And the two systems still differ in architecture, capacity, and training regime, so the comparison isolates input modality but not the causal effect of freezing a pretrained encoder per se.

5.4 Architecture-Matched Controls

The corruption comparison leaves open whether NDB-JEPA's advantages come from the pretrained representation or from the vision-transformer architecture. We test this with three controls trained on RGB input with the same two-tower head (Table 3): (i) the I-JEPA ViT-H/14 architecture with random, frozen weights; (ii) a ViT-B/16 trained end-to-end from scratch, for 20 and 100 epochs; and (iii) the same ViT-B/16 fine-tuned end-to-end from supervised ImageNet weights. End-to-end ViT-H/14 does not fit in 12 GB. All controls use batch size 32 and a head learning rate of 5×10^{-4}, with encoder learning rates of 10^{-4} from scratch and 10^{-5} from supervised initialization.

Table 3: Architecture-matched controls: Top-1 rectangle success (%) at IoU > 0.25, mean $\pm$ standard deviation over three seeds. All systems use RGB input, and all ViT systems use the same two-tower head. ViT-H is ViT-H/14 and ViT-B is ViT-B/16; "frozen" trains only the head, "e2e" trains all parameters end-to-end, and "ep." is epochs. The clean columns train on 100% or 10% of the training data; the Cornell-C columns evaluate the 100% models.

	Clean		Cornell-C	
System	100%	10%	Dark	Occlusion
NDB-JEPA (I-JEPA ViT-H, frozen)	86.9 ± 1.3	73.8 ± 5.8	81.3 ± 3.4	76.0 ± 3.4
Random-init ViT-H, frozen	60.7 ± 6.8	31.8 ± 4.7	0.0 ± 0.0	25.5 ± 2.8
ViT-B from scratch, e2e, 20 ep.	40.4 ± 7.9	22.1 ± 19.3	3.7 ± 3.9	26.2 ± 8.4
ViT-B from scratch, e2e, 100 ep.	50.6 ± 4.1	39.0 ± 2.3	2.6 ± 2.6	31.5 ± 3.0
ViT-B supervised init, e2e	68.2 ± 2.6	39.0 ± 5.1	63.3 ± 2.3	37.5 ± 15.7
GR-ConvNet (RGB), e2e	82.0 ± 5.8	36.0 ± 4.1	8.2 ± 14.3	35.2 ± 28.8

The controls attribute NDB-JEPA's advantages to pretraining rather than to the architecture. With the same ViT-H/14 and head, random frozen weights reach only 60.7% on clean images and 31.8% with 10% of the data, against 86.9% and 73.8% for I-JEPA, and fail completely under darkening (0.0%). A ViT-B/16 trained from scratch does no better: even after 100 epochs it reaches 50.6% on clean images and collapses under darkening (2.6%), in some runs proposing no grasp for most test images. The fine-tuned supervised ViT-B/16 retains 63.3% under darkening but reaches only 39.0% with 10% of the data, so conventional pretraining with fine-tuning does not reproduce the label efficiency of the frozen I-JEPA features. Occlusion robustness appears only with frozen I-JEPA (76.0%, against at most 37.5% for every other system). The end-to-end controls are smaller than ViT-H/14, but the random-init ViT-H/14 control matches NDB-JEPA's architecture exactly; the controls do not, however, separate freezing from fine-tuning.

6 Conclusion

We presented NDB-JEPA, a grasp point detector that trains lightweight confidence and geometry heads on the frozen patch embeddings of a large I-JEPA encoder. In controlled comparisons on the Cornell Grasping Dataset, I-JEPA provided the strongest representation among the four pretraining paradigms tested at matched scale, reaching 86.9% Top-1 rectangle success, within 1.5 points of an end-to-end GR-ConvNet baseline. The frozen predictive representation is markedly more label-efficient, reaching approximately 74% success with only 10% of the training labels versus approximately 40% for GR-ConvNet trained from scratch, and it is substantially more robust to darkening and occlusion; once the baseline is retrained on RGB only, so that both systems observe the same corrupted signal, NDB-JEPA equals or exceeds it on every corruption tested except Gaussian noise. Supervised ViT and DINOv2 features underperform the from-scratch baseline in low-data regimes, so the frozen backbone's pretraining objective matters. Overall, our results support frozen joint-embedding predictive representations as a practical, data-efficient basis for grasp perception in settings where collecting exhaustive grasp annotations is prohibitively expensive.

6.1 Limitations

Fine-tuning the backbone, even with parameter-efficient methods such as LoRA, was infeasible on our single GPU hardware setup; all adaptation therefore happens downstream of the frozen encoder. Inference is also slower (61.4 ms versus 6.3 ms per image for GR-ConvNet at batch size 1). Our default GR-ConvNet baseline uses RGB-D while NDB-JEPA uses RGB; we address this confound directly by additionally training an RGB-only GR-ConvNet for the corruption study, but the two systems still differ in architecture and training regime, so the comparison isolates input modality rather than the causal effect of freezing a pretrained encoder. Our image-wise Cornell split does not establish generalization to unseen objects or other datasets, and we did not test an ImageNet-pretrained CNN or parameter-efficient adaptation.

AI USE STATEMENT

In this work, we used generative AI tools, specifically a large-language-model coding and writing assistant, for the following tasks with required disclosure: assistance in writing parts of the code base, most notably the corruption suite experiment. An LLM was used to edit figures, catch grammar mistakes, make minor edits to improve writing flow and conciseness, and remove redundancies and move some information to the appendix to reach the paper page limit. An LLM was also used to find some of the relevant papers for the literature review. We have not used generative AI tools to generate data sets (Cornell-C images are produced by deterministic image transformations in our code), to develop theoretical models, to formulate or prove mathematical claims, or for translation, and the remaining required-disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for tasks with recommended disclosure: creating and editing software code, modifying figures, identifying relevant literature and formatting references, and editing the manuscript text. We have reviewed all AI-assisted work. Every reported number was produced by running our code on the datasets described above, and AI-written experiment code was reviewed by the authors before it was run. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses the publicly available Cornell Grasping Dataset, which contains images of household objects and no human subjects or personally identifying information. All pretrained encoders are publicly released checkpoints used under their respective licenses. We foresee no direct harmful applications of grasp detection beyond those already inherent to robotic manipulation, and we report the limitations of our evaluation protocol in Section 6.1.

REPRODUCIBILITY STATEMENT

The evaluation protocol, data split, and Top-1 rectangle metric are specified in Section 4.2, and the grasp-head topology, loss, and optimization settings in Section 4. The Cornell-C corruption parameters and their deterministic seeding scheme are described in Section 4.6, and the nested low-shot subsampling procedure in Section 4.5. The backbone comparison, low-shot, corruption, and architecture-control results are means over seeds $\{0, 1, 2\}$ produced by a single shared evaluation implementation, and the settings of the architecture-matched controls are given in Section 5.4. The system ablations use the separate protocol stated in Appendix A; supplementary figures for them are given in Appendix F, and their validation-split re-run in Appendix B. Source code, configuration files, and per-run result files are provided as anonymized supplementary material.

REFERENCES

M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture, 2023. URL https://arxiv.org/abs/2301.08243.

K. Bousmalis, A. Irpan, P. Wohlhart, Y. Bai, M. Kelcey, M. Kalakrishnan, L. Downs, J. Ibarz, P. Pastor, K. Konolige, S. Levine, and V. Vanhoucke. Using Simulation and Domain Adaptation to Improve Efficiency of Deep Robotic Grasping. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4243–4250, 2018.

N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-End Object Detection with Transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

A. Depierre, E. Dellandréa, and L. Chen. Jacquard: A Large Scale Dataset for Robotic Grasp Detection. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018.

A. Dosovitskiy, G. Ros, F. Codevilla, A. López, and V. Koltun. CARLA: An Open Urban Driving Simulator. In *Proceedings of the Conference on Robot Learning (CoRL)*, pp. 1–16, 2017.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. ImageNet-trained CNNs are Biased Towards Texture; Increasing Shape Bias Improves Accuracy and Robustness. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked Autoencoders Are Scalable Vision Learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, 2022.

D. Hendrycks and T. Dietterich. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

Y. Jiang, S. Moseson, and A. Saxena. Efficient Grasping from RGBD Images: Learning Using a New Rectangle Representation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3304–3311. IEEE, 2011.

S. Kumra, S. Joshi, and F. Sahin. Antipodal Robotic Grasping using Generative Residual Convolutional Neural Network. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 9626–9633, 2020.

Y. LeCun. A Path Towards Autonomous Machine Intelligence, 2022. Version 0.9.2.

I. Lenz, H. Lee, and A. Saxena. Deep learning for detecting robotic grasps. *The International Journal of Robotics Research*, 34(4-5):705–724, 2015.

J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. Aparicio Ojea, and K. Goldberg. Dex-Net 2.0: Deep Learning to Plan Robust Grasps with Synthetic Point Clouds and Analytic Grasp Metrics. In *Proceedings of Robotics: Science and Systems (RSS)*, 2017.

A. Majumdar, K. Yadav, S. Arnaud, Y. J. Ma, C. Chen, S. Silwal, A. Jain, V.-P. Berges, P. Abbeel, J. Malik, D. Batra, Y. Lin, O. Maksymets, A. Rajeswaran, and F. Meier. Where are We in the Search for an Artificial Visual Cortex for Embodied Intelligence? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Meta AI. facebook/ijepa_vith14_1k: I-JEPA Model (Huge, fine-tuned on ImageNet-1K). `https://huggingface.co/facebook/ijepa_vith14_1k`, 2023. Hugging Face model repository.

D. Morrison, P. Corke, and J. Leitner. Learning robust, real-time, reactive robotic grasping. *The International Journal of Robotics Research*, 39(2-3):183–201, 2020.

S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta. R3M: A Universal Visual Representation for Robot Manipulation. In *Proceedings of the Conference on Robot Learning (CoRL)*, pp. 892–909, 2022.

M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jégou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research (TMLR)*, 2024.

I. Radosavovic, T. Xiao, S. James, P. Abbeel, J. Malik, and T. Darrell. Real-World Robot Learning with Masked Visual Pre-training. In *Proceedings of the Conference on Robot Learning (CoRL)*, pp. 416–426, 2023.

J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel. Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 23–30, 2017.

S. Wang, Z. Zhou, and Z. Kan. When Transformer Meets Robotic Grasping: Exploits Context for Efficient Grasp Detection. *IEEE Robotics and Automation Letters*, 7(3):8170–8177, 2022.

Table 4: Loss comparison: average IoU and grasp-success rate on the Cornell test set under the ablation protocol. MSE with smoothing (radius 4) performs significantly worse.

Loss	IoU	Grasp Success (%)
BCE	0.466	85.4
MSE	0.469	**87.6**
MSE w/ Smoothing	0.372	61.8

A System Ablations

The experiments in this section guided the design of NDB-JEPA and were conducted prior to baseline comparisons. They use the I-JEPA system (with an upsampled 32×32 anchor grid unless stated otherwise) and are evaluated under a more permissive protocol: we report the average IoU of the top-ranked predictions, and count a grasp as successful if its IoU exceeds 0.25 and its angular error is below 25° (Mahler et al., 2017). Their absolute values are therefore not comparable to the Top-1 results above; they are reported to justify design choices rather than as headline performance. These ablations were scored on the test split before the configuration was fixed. Of the choices they informed, only the MSE confidence loss carries into the backbone, low-shot, and corruption experiments, and it is also the loss selected by the grid search below; those experiments use neither upsampling nor, for any system including GR-ConvNet, data augmentation. Appendix B repeats the loss, upsampling, and augmentation ablations on the validation split under the canonical metric with three seeds.

A.1 Hyperparameter Search

We swept learning rates $\{10^{-2}, 5 \times 10^{-3}, 10^{-4}, 5 \times 10^{-4}\}$, batch sizes $\{16, 32, 64, 128\}$, and confidence losses {BCE, MSE}, training each of the 32 configurations for 10 epochs on a 5% subset of Jacquard (2,724 images) and ranking them by validation loss. We concluded on a stable near-optimal setting (learning rate 5×10^{-4}, batch size 128, MSE loss) whose training and validation losses decrease smoothly with only a small gap between them.

A.2 Loss Functions

We compare three confidence losses (Table 4). Standard BCE and MSE perform within the margin of error of one another and select identical top-5 grasps: because the geometry terms are heavily down-weighted and the confidence targets are binary, both losses converge to the same solution, with BCE producing near-binary confidences and MSE softer ones concentrated on the object. MSE with spatial smoothing performs significantly worse; it spreads confidence onto patches that merely neighbor the object, which does not translate into valid grasps.

A.3 Upsampling

Because predicted grasps are anchored at patch centers, the anchor grid discretizes a continuous localization problem. Bilinear upsampling of the final encoder output doubles the grid resolution, quadrupling the number of anchors. Table 5 shows that upsampling at inference time performs best regardless of whether the head was trained with it, while training with upsampling but predicting without it hurts, indicating that upsampling also affects what the grasp head learns.

A.4 Data Augmentation

The Cornell dataset has a glaring bias: grasped items tend to be centered in the image, so the ViT positional embeddings could bias the grasp head toward central anchors. We test this with 180-degree rotations and random translations. As Table 6 shows, the model trained with augmentation retains a success rate comparable to non-augmented accuracies elsewhere (about 88%), while the model trained without augmentation largely fails on the transformed test set; its confidence maps point to where the object would have been before augmentation, indicating positional-embedding

Table 5: Upsampling ablation under the ablation protocol. Models with upsampling at inference time perform best.

Training : Runtime	IoU	Grasp Success (%)
None : None	0.469	87.6
None : Upsample	0.526	**92.1**
Upsample : None	0.456	77.5
Upsample : Upsample	**0.539**	87.6

Table 6: Effect of training-time augmentation under the ablation protocol. Both models use the grid-searched hyperparameters with upsampling, and both are evaluated on a test set augmented at rate 0.25.

Data Augmentation	IoU	Grasp Success (%)
None	0.393	61.8
Rate $= 0.25$	0.497	**88.8**

bias. The non-augmented model has still clearly learned to predict grasps, but augmentation is vital for generalizing the grasp head.

B Validation-Split Ablations

The system ablations of Appendix A were each run once and scored on the test split under the ablation protocol. To keep design decisions independent of the test set, we repeat the loss, upsampling, and augmentation ablations on the validation split under the canonical metric of Section 4.2, with seeds $\{0, 1, 2\}$. Each variant is the I-JEPA system trained for 20 epochs at batch size 32 and a learning rate of 5×10^{-4}, with the checkpoint selected by validation loss, and changes one factor at a time. The hyperparameter search is not repeated.

B.1 Loss Functions

We compare the same three confidence losses (Table 7): BCE with a positive-class weight of 40, MSE, and MSE against spatially smoothed targets, whose confidence decays linearly with the Chebyshev distance to the nearest grasp-center cell and reaches zero at a radius of 4 cells.

Table 7: Loss comparison on the Cornell validation split under the canonical metric: mean IoU and Top-1 rectangle success at IoU > 0.25 and IoU > 0.50, mean $\pm$ standard deviation over three seeds.

Loss	Mean IoU	@0.25	@0.50
BCE	0.454 ± 0.002	86.0 ± 1.7	34.8 ± 1.7
MSE	0.458 ± 0.005	85.6 ± 1.3	34.5 ± 1.3
MSE w/ Smoothing	0.431 ± 0.012	82.6 ± 4.0	28.0 ± 0.7

BCE and MSE again perform within the margin of error of one another (86.0% versus 85.6%). Smoothed targets are only modestly worse ($82.6 \pm 4.0\%$), with lower mean IoU and lower strict-threshold success (28.0% versus 34.5%), consistent with confidence spreading onto patches that merely neighbor a grasp center. The much larger drop in Table 4 (61.8%) does not reproduce. In that run the geometry mask M_{uv} also admitted neighboring cells whose smoothed confidence exceeds 0.5 and whose geometry targets are all zero; for this re-run we restrict the mask to grasp-center cells, which leaves binary targets unchanged.

B.2 UPSAMPLING

We vary whether bilinear upsampling is applied during training and at inference (Table 8).

Table 8: Upsampling ablation on the Cornell validation split under the canonical metric, mean $\pm$ standard deviation over three seeds.

Training : Runtime	Mean IoU	@0.25	@0.50
None : None	0.458 ± 0.005	85.6 ± 1.3	34.5 ± 1.3
None : Upsample	0.491 ± 0.007	86.0 ± 0.7	52.7 ± 2.9
Upsample : None	0.446 ± 0.011	83.7 ± 2.6	32.2 ± 1.7
Upsample : Upsample	0.505 ± 0.012	88.3 ± 2.4	54.5 ± 1.1

Upsampling during both training and inference performs best at both thresholds (88.3% and 54.5%). Its effect is mainly on localization precision. At $\mathrm{IoU} > 0.25$ the gain over the native grid is 2.7 points, comparable to the seed-to-seed variation, whereas at $\mathrm{IoU} > 0.50$ it is 20.1 points, in line with the higher oracle ceiling of the finer grid (Appendix E). Upsampling only at inference recovers most of the strict-threshold gain (52.7%), and upsampling only during training slightly lowers success (83.7%). The ordering in Table 5, where upsampling only at inference performed best (92.1%), does not reproduce.

B.3 DATA AUGMENTATION

We apply the same 180-degree rotations and random translations, each with probability 0.25 per image and with translations of up to 100 pixels in each axis and wrap-around, to the training split only. We evaluate on the clean validation split and on a copy of it transformed at the same rate, with the ground-truth rectangles transformed accordingly (Table 9).

Table 9: Effect of training-time augmentation on the Cornell validation split: Top-1 rectangle success (%) at $\mathrm{IoU} > 0.25$ under the canonical metric, mean $\pm$ standard deviation over three seeds. The transformed split applies the training transformations at rate 0.25.

Data Augmentation	Clean val	Transformed val
None	85.6 ± 1.3	84.5 ± 1.3
Rate $= 0.25$	85.6 ± 0.7	85.2 ± 2.0

Augmentation has no measurable effect. Both models reach 85.6% on the clean validation split, and on the transformed split the model trained without augmentation reaches 84.5% against 85.2% with augmentation, well within the seed-to-seed variation. We therefore find no evidence of the positional-embedding bias suggested by Table 6. The earlier ablation applied augmentation once, when the dataset was constructed, to every split, and its result does not reproduce under this protocol.

C COMPARISON WITH PUBLISHED CORNELL RESULTS

Absolute values under this protocol are lower than the numbers usually quoted for Cornell, where the strongest published detectors report 97–98% success (Kumra et al., 2020; Wang et al., 2022). The difference is one of protocol rather than implementation. Published results are typically obtained with five-fold cross-validation over this small dataset, heavy train-time augmentation, and post-processing of the predicted quality map before the peak grasp is selected, each of which raises reported success substantially. We instead use a single fixed image-wise split, hold the decoding and selection procedure identical across every system, and apply no test-time processing, because our claims are comparative rather than absolute. The quantity of interest throughout is the difference between systems measured under one protocol, and every number we report is produced by the same evaluation code; it is important that our results should not be read as a leaderboard entry against separately tuned published pipelines.

D A Hypothesis for the Backbone Ordering

We can offer a hypothesis for this ordering, though our experiments measure outcomes rather than mechanisms. Grasp prediction is a dense task: the head must resolve, at each spatial location, whether that patch admits a stable grip and at what orientation. I-JEPA's objective requires predicting the *representations* of masked regions from visible context, which rewards encoding the structure and extent of objects at patch granularity. MAE's pixel-reconstruction objective also preserves spatial detail, which is consistent with it placing second, but it allocates capacity to appearance detail such as texture and color that grasp geometry does not require. DINOv2's image-level self-distillation optimizes a global objective for which patch embeddings need only be discriminative in aggregate, so precise per-patch spatial information may be less well preserved. This is a plausible explanation for why its larger encoder does not translate into better grasp features. Testing this would require probing the spatial content of each representation directly, which we leave to future work.

E Upper Bound Imposed by the Anchor Grid

Finally, because of the 16×16 anchor grid, we solve a continuous localization problem in a discretized space, which places an upper bound on the achievable IoU. We quantify this bound with an oracle that emits the exact ground-truth angle and opening but is constrained to place the grasp center at an anchor-cell center. On the Cornell test split this oracle attains a mean best IoU of 0.56 on the 16×16 grid: it reaches 100% success at the standard $\text{IoU} > 0.25$ threshold, but only 64.0% at the stricter $\text{IoU} > 0.50$. Upsampling to 32×32 raises the mean oracle IoU to 0.70 and the strict-threshold ceiling to 89.9%. The anchor grid is therefore not the binding constraint at the standard operating point, but it dominates at $\text{IoU} > 0.50$, where reported numbers should be read against this ceiling rather than against 100%.

F Supplementary Figures

This appendix collects supplementary figures for the system ablations of Appendix A. Figures 4 and 5 show training and validation loss curves for two configurations from the hyperparameter search, labeled C02 and C31; C31 is the selected setting. Figure 6 compares the confidence maps produced by the three confidence losses, Figure 7 shows predictions with and without upsampling, and Figure 8 shows predictions on augmented test images from models trained with and without augmentation.

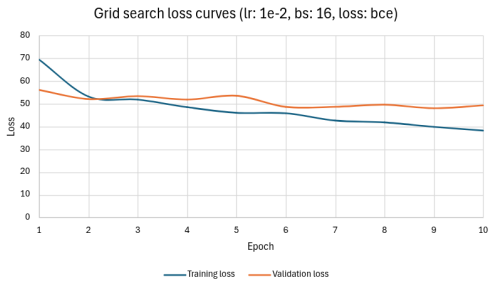


Figure 4: Although C02 achieved a lower validation loss, the training loss continues decreasing while the validation loss remains flatter and more oscillatory, suggesting weaker generalization.

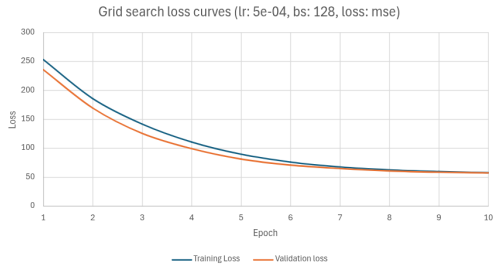


Figure 5: C31 training and validation loss decrease smoothly, with only a small gap between them. This suggests stable optimization and limited overfitting.

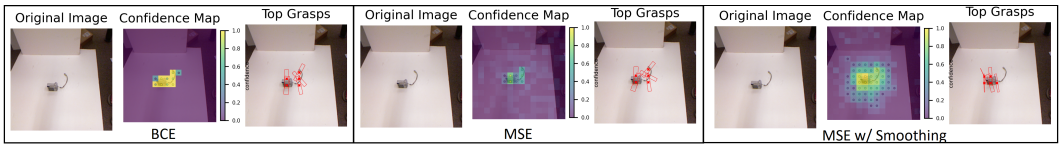


Figure 6: BCE Loss (left): distinct confidences either near 0 or 1; MSE Loss (middle): less confident overall but similar top grasps to BCE; MSE with Smooth Loss (right): Much of the confident anchors fall outside the object, but predictions are all near the densest labeled area.

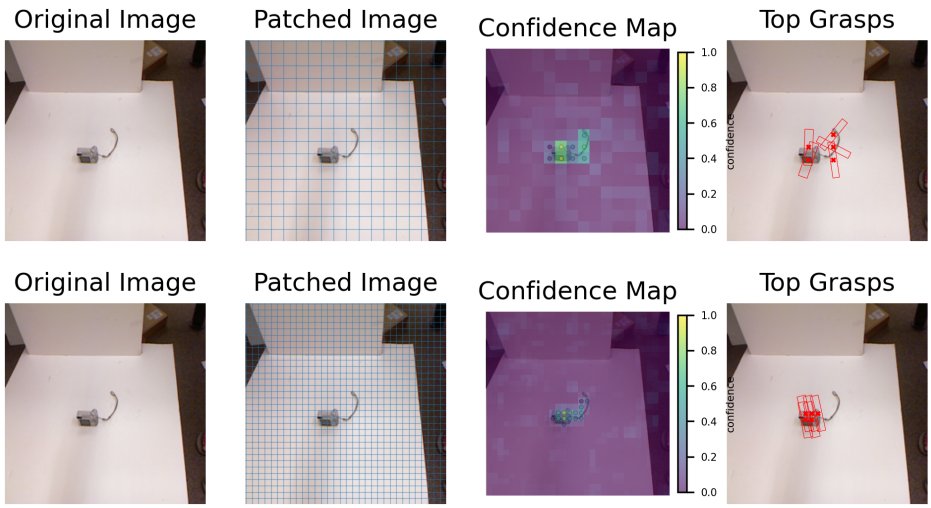


Figure 7: The top four images correspond to no upsampling. The bottom four images have bilinear upsampling so that there are 32x32 images patches; this leads to denser anchor point positions for predicted grasps which in this case leads to none of the grasp center points fall outside the body of the camera unlike with the non-upsampled example.

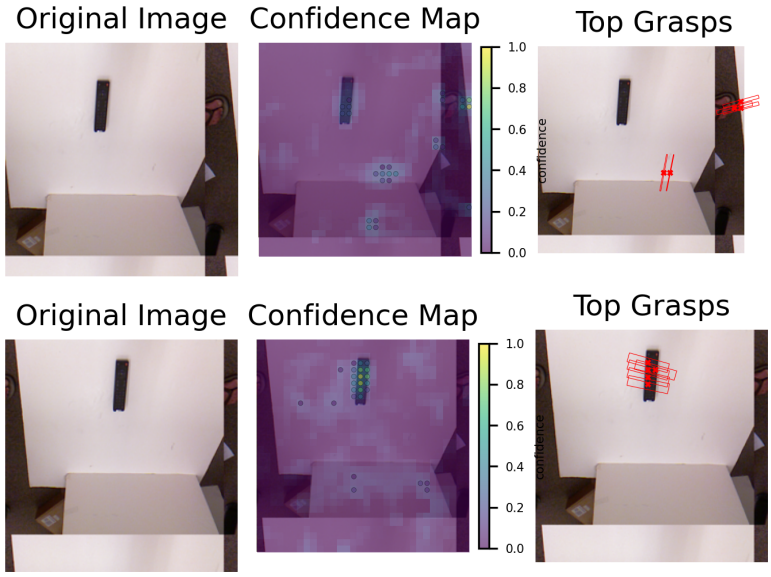


Figure 8: Model trained with no data augmentation (top) and Model trained with data augmentation of 25% (bottom): Both models were evaluated on an augmented dataset.